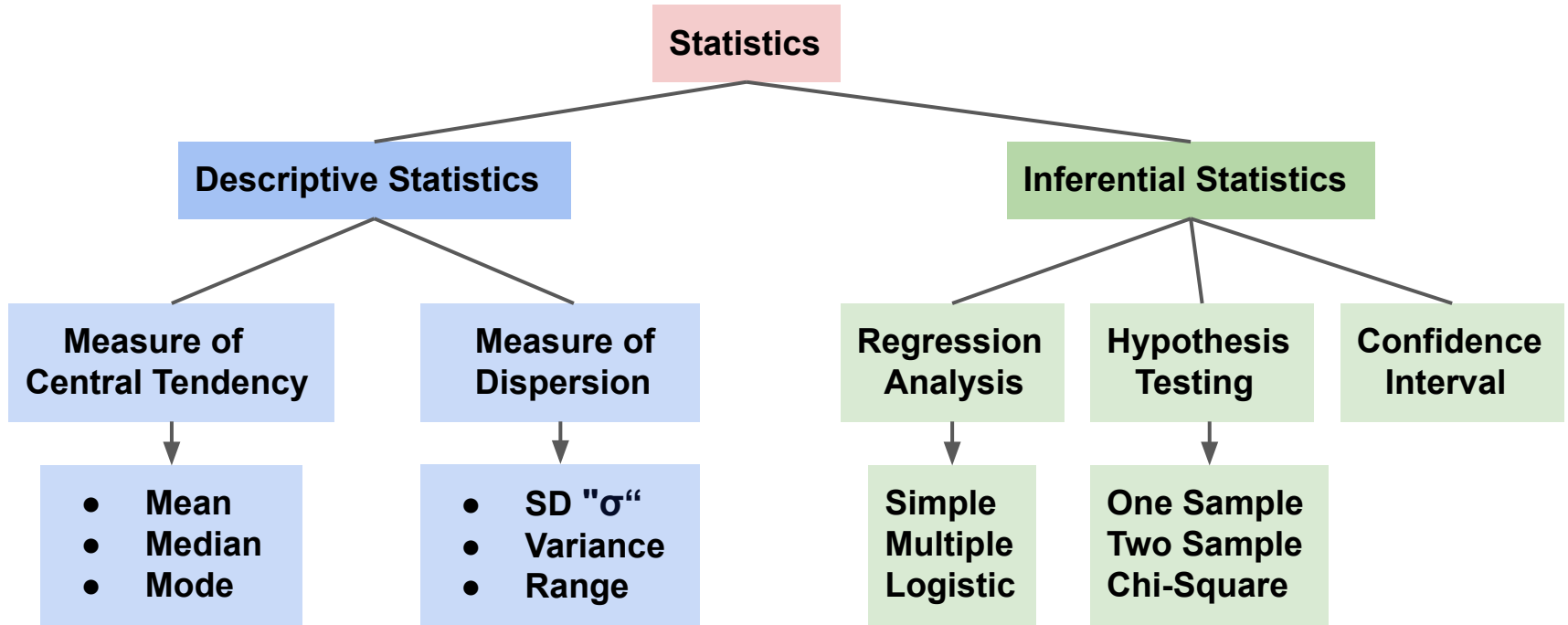


Statistics

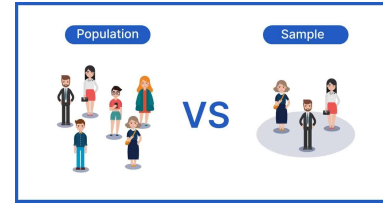
Introduction to Statistics

- The field of statistics is the practice and study of collecting and analyzing data.
- Main Types:
 1. Descriptive or summary statistics
 - They are used to describe or summarize our data either through charts/graphs or through numerical calculations using measures of central tendency.
 - Here conclusion are drawn based on already known data.
 2. Inferential Statistics
 - Involves making inferences or predictions about a population based on a sample of data.
 - Uses probability theory and statistical techniques to draw conclusions about a larger group.

Hierarchy



Population vs Sample



- **Size**
 - The **population** represents the complete set, while **sample** is a smaller subset taken from the population.
- **Representation**
 - The **population** includes all the individuals or elements of interest, while the **sample** represents a subset that is selected to represent the **population**.
- **Feasibility**
 - It is often impractical or impossible to collect data from the entire **population**, so a **sample** is taken to make data collection and analysis more manageable.
- **Generalizability**
 - The goal of sampling is to obtain a representative sample that accurately reflects the **population**. Statistical analysis is then used to make inferences and generalize the findings from the **sample** to the larger **population**.
- **Variation**
 - The **population** may exhibit greater variability since it includes all possible values. The **sample** may have variability as well but to a lesser extent due to its smaller size.

What Statistics Can do?

- **Stock Market Data Analysis**

- Stats techniques like moving averages, regression analysis, and correlation analysis are used to analyze historical stock prices, predict future price movements, and identify relationships between different stocks or market indices.

- **Weather Forecasting**

- Statistical methods such as regression analysis, time series analysis, and probability distributions are used to analyze temperature, humidity, wind patterns, and other meteorological variables to forecast weather conditions accurately.

- **Natural Disaster Prediction**

- Statistical techniques like cluster analysis, spatial analysis, and trend analysis can be used to identify areas prone to earthquake or regions susceptible to flooding, etc to prevent from possible disaster.

- **Sports**

- Advanced statistical techniques like machine learning algorithms are used to analyze large volumes of sports data for predictive modeling.

Many more....

Limitations of Statistics

- **Sample Bias**

- Stats relies on samples to make inferences about population.
- If the sample used is not truly representative of the population of interest, it can introduce bias and lead to inaccurate conclusions.

- **Measurement Errors**

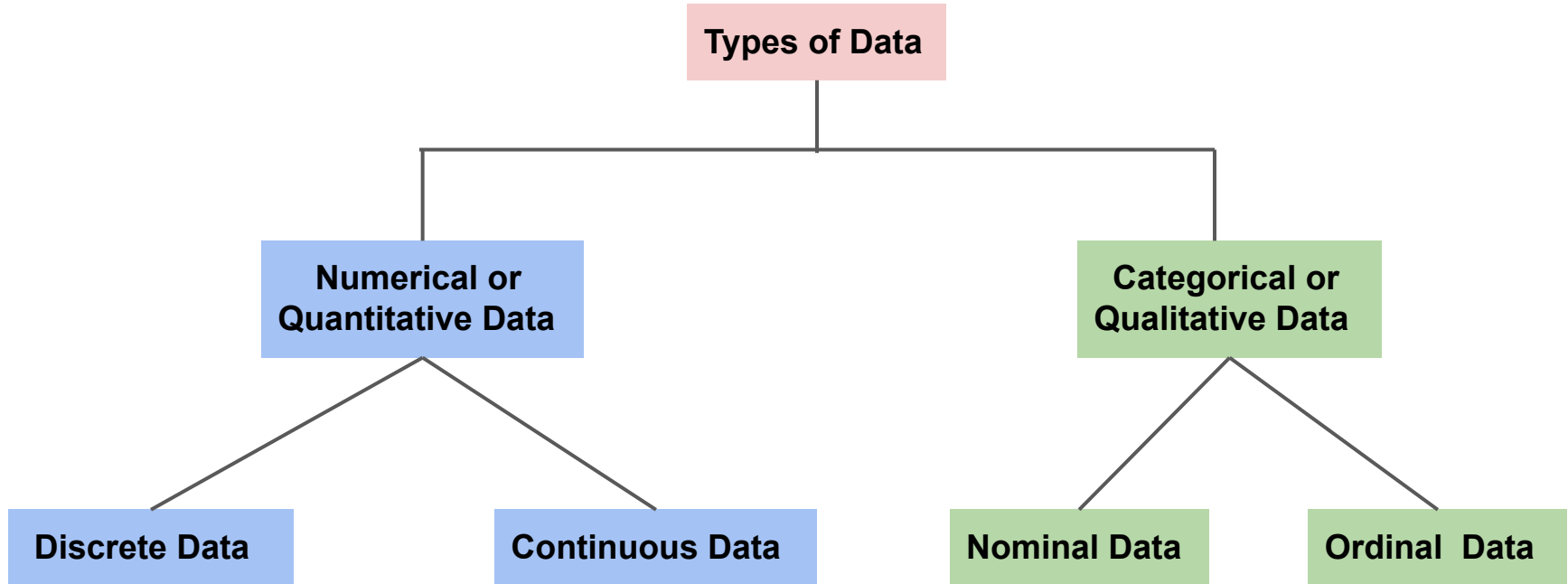
- Measurement errors can occur due to various factors such as human error, equipment limitations, or data entry mistakes. Statistical analysis from such data could be error prone.

- **Limitation in Explaining Causality**

- Example:
 - Statistics can tell us if rock music is more popular than jazz, based on total sales, or whether women live longer than men.
 - However, we can't use statistics to find out why relationship exists, such as why people like different types of music or why women live longer than men.

and many more

Types of Data in Statistics



Measure of Central Tendency

- Also called as Average / Summary Statistics.
- Measure of Central Tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- Measure of Central Tendency are:
 - Mean
 - Median
 - Mode

Mean

- The mean is calculated by summing up all the values in a dataset and dividing it by total number of values.
- **Mean = Sum of all values / Count of all values**
- **Example:**
 - Compute average income of group of individuals.
 - Income = [10000, 20000, 25000, 5000, 18000, 30000, 27000]
 - Mean (Income) = $(10000+20000+25000+5000+18000+30000+27000) / 7$
 $= 19285.71$

Median

- Median is the middle value in an ordered dataset, separating it into two equal halves.
- Before computing median, we need to arrange our data in an ascending order.
- There are 2 cases in median computation:
 - a. Compute median when number of samples are even
 - b. Compute median when number of samples are odd

Median (Even Number of Samples)

- Consider dataset: **12, 18, 20, 24, 30, 36**
- **Step 1:** Sort the dataset in ascending order: **12, 18, 20, 24, 30, 36**
- **Step 2:** For even number of samples, take the average of the two middle values.
 - The two middle values are 20 and 24.
 - Median = $(20 + 24) / 2 = 22$
- Therefore, **Median = 22**

Median (Odd Number of Samples)

- Consider dataset: **10, 15, 18, 22, 25, 30, 36**
- **Step 1:** Sort the dataset in ascending order: **10, 15, 18, 22, 25, 30, 36**
- **Step 2:** For odd number of samples, the median is the middle value.
 - Median = $(n + 1) / 2$ th item
= 4th item
- Therefore, **Median = 22**

Mode

- Mode is the most frequent value in a dataset.
- Mode is generally suitable for Categorical Data.
- **Example:**
 - Dataset: [10, 20, 20, 30, 20, 40]
 - Mode = 20

Q. Why Median and Mode, although there is Mean?

Consider dataset: [3, 3, 3, 3, 3, 100]

1. Arithmetic Mean

- Mean = $115 / 6 = 19.167$

2. Median

- Median = $(3 + 3) / 2 = 3$

3. Mode

- Mode = 3

- **Arithmetic mean = 19.167**, this means mean is affected by large number 100. Due to single large number mean is shifted towards the large number, which leads to false representative of the given dataset
- **Median = 3**, Median is unaffected by large numbers i.e 100, which may be outliers.
- **Mode = 3**, Mode is also unaffected by the large numbers, which may be an outlier.

Summary (Mean, Median, Mode)

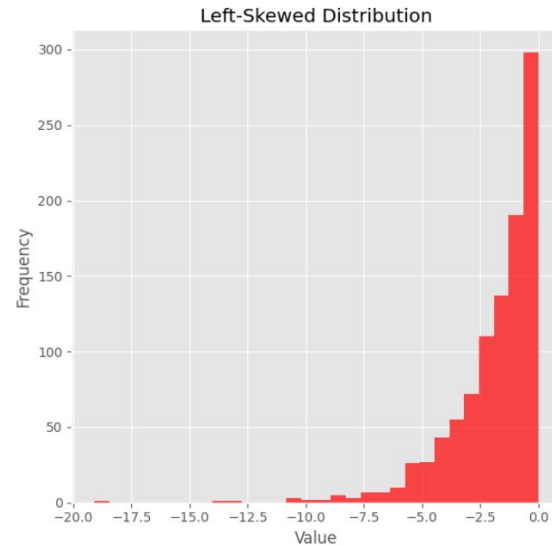
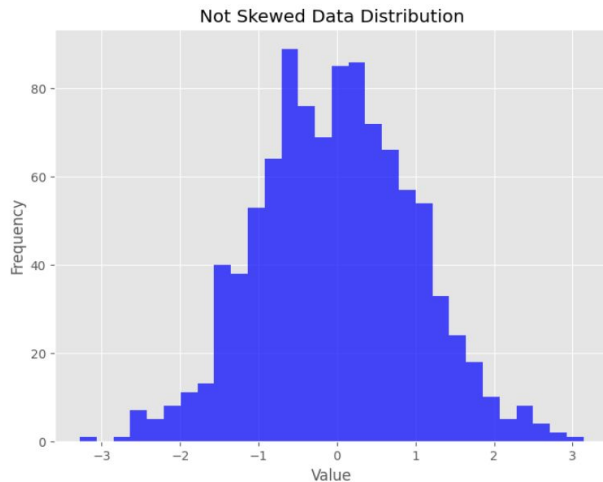
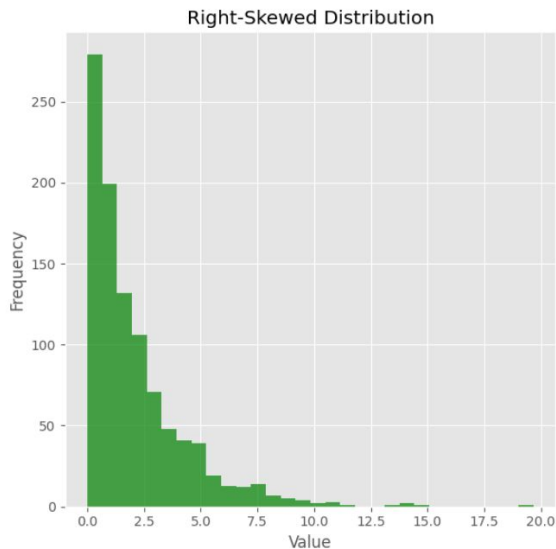
- **Arithmetic Mean** is more sensitive to outliers.
- **Mode** and **Median** are commonly unaffected by the outliers.
- Normally, The mode is used for categorical data where we wish to know which is the most common category.
- One of the problem with mode is that, it is not unique, so it leaves problem when we have two or more values that share the highest frequency.

$$MODE = 3 MEDIAN - 2 MEAN$$

When to use (Mean, Median, Mode)

Type of Variable	Best Measure of Central Tendency
Nominal	Mode
Ordinal	Median
Interval / Ratio (Not Skewed)	Mean
Interval / Ratio (Skewed)	Median

Not Skewed vs Right Skewed vs Left Skewed



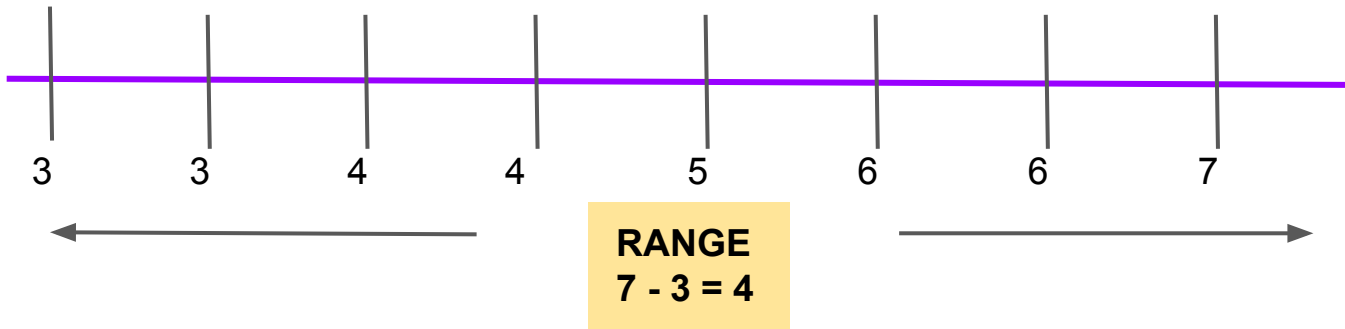
Measure of Dispersion

- Dispersion means spread.
- How far is the data from the central tendency (mean, median, mode) ?
- We'll cover:
 - Range
 - Standard Deviation
 - Variance

Measure of Dispersion (Range)

- Range is defined as the difference between the maximum and the minimum observation of the given data.

- **Range = $X_{\max} - X_{\min}$**



- Poor measure of dispersion.
- Do not give a good picture of the overall spread of the data with respect of central tendency.

Measure of Dispersion (Variance & SD)

- Consider population data as:

Population Data 1	Population Data 2
<ul style="list-style-type: none">[2, 2, 3, 3]	<ul style="list-style-type: none">[0, 0, 5, 5]
<ul style="list-style-type: none">Mean = $(2 + 2 + 3 + 3) / 4$ = 2.5	<ul style="list-style-type: none">Mean = $(0 + 0 + 5 + 5) / 4$ = 2.5
<ul style="list-style-type: none">Data points are close to mean	<ul style="list-style-type: none">Data points are far from mean

- Here mean are same for 2 data distributions
- How do we say the 2 distributions are different?**
 - Hint: Use Variance and Standard Deviation

Variance and SD Calculation

i	xi	mean	xi - mean	(xi - mean)2
1	2	2.5	-0.5	0.25
2	2	2.5	-0.5	0.25
3	3	2.5	0.5	0.25
4	3	2.5	0.5	0.25

$$\text{Population Variance}(\sigma^2) = \frac{\sum(x_i - \mu)^2}{N}$$

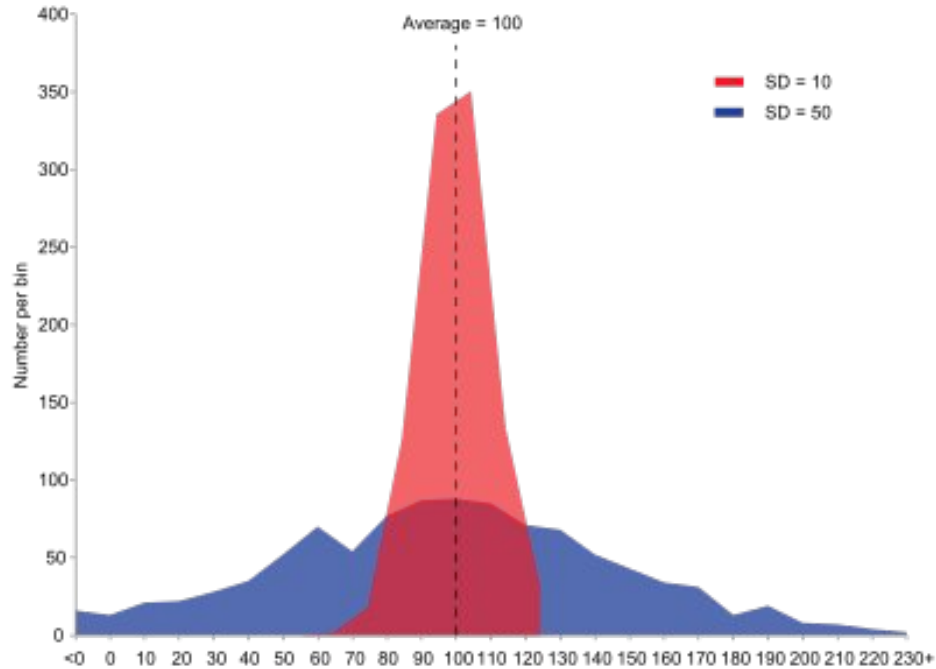
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

- Table1
 - variance = 0.25
- Table 2
 - variance = 6.25

i	xi	mean	xi - mean	(xi - mean)2
1	0	-2.5	-2.5	6.25
2	0	2.5	-2.5	6.25
3	5	2.5	2.5	6.25
4	5	2.5	2.5	6.25

“Dataset in table 2 are more dispersed than in table 1”

Low vs High Standard Deviation



[Source](#)

Notation (Population vs Sample)

Parameter	Population Notation	Sample Notation	Description
Mean	μ	\bar{x}	Average of all values in the population or sample.
Variance	σ^2	s^2	Measure of dispersion of values around the mean.
Standard Deviation	σ	s	Square root of the variance; indicates data spread.
Size/Count	N	n	Total number of items in the population or sample.
Data Point	x_i	x_i	A single observation from the population or sample.
Sum	$\sum x_i$	$\sum x_i$	Sum of all data points in the population or sample.

Frequency Distribution

- Data is generally classified as Numerical and Categorical.
- Frequency Distributions help to understand the distribution or pattern of data by displaying counts or frequency of various values or intervals.
- **Steps**
 - Data Collection
 - Data Sorting
 - Identify Categories (For Categorical Data)
 - Create Intervals (For Numerical Data)
 - Count Frequencies
 - Create Frequency Table
 - Visual Representation:
 - Bar Chart or Histograms

Frequency Distribution (Categorical Data)

- Consider Categorical Data:

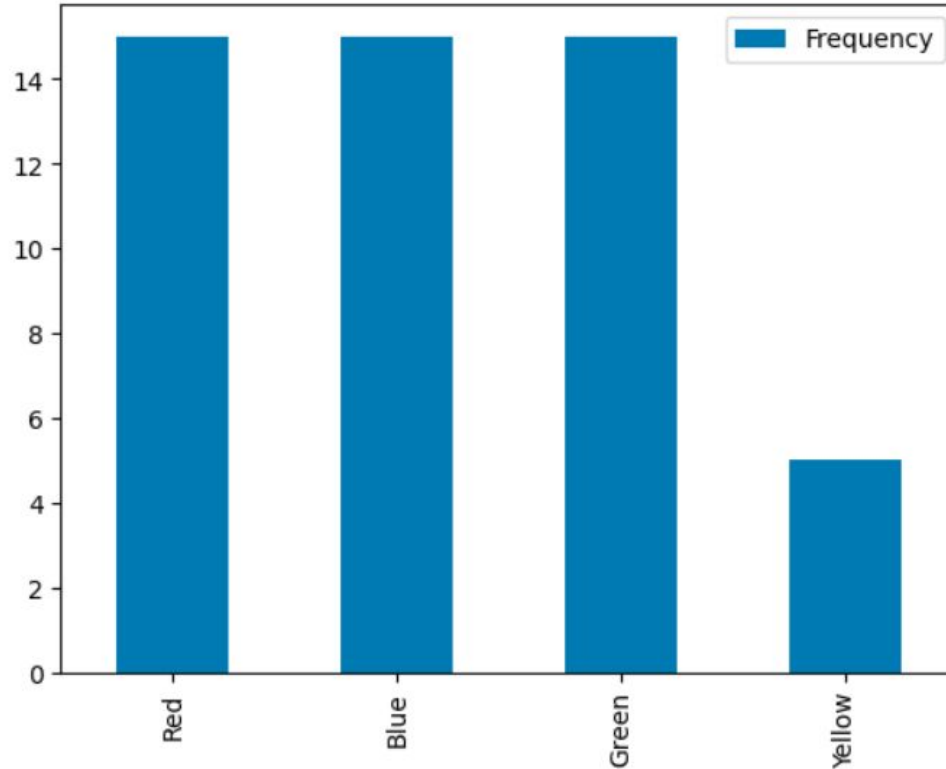
- ```
colors = [
 "Red", "Blue", "Green", "Red", "Yellow", "Blue", "Green", "Green", "Blue", "Red",
 "Red", "Blue", "Yellow", "Green", "Green", "Blue", "Red", "Red", "Red", "Blue",
 "Green", "Yellow", "Blue", "Blue", "Green", "Red", "Red", "Green", "Blue", "Green",
 "Yellow", "Red", "Blue", "Green", "Blue", "Red", "Blue", "Green", "Blue", "Red",
 "Green", "Green", "Yellow", "Red", "Blue", "Green", "Blue", "Red", "Green", "Red"
]
```

- Construct Frequency Table:

| Color  | Frequency |
|--------|-----------|
| Red    | 15        |
| Blue   | 15        |
| Green  | 15        |
| Yellow | 5         |

# Frequency Distribution (Categorical Data)

- Construct Bar Graph



# Frequency Distribution (Numerical Data)

- **Consider Dataset:**

```
ages = [
 28, 34, 25, 28, 22, 31, 25, 27, 30, 24,
 22, 29, 26, 31, 28, 34, 30, 25, 27, 26,
 23, 29, 32, 28, 26, 24, 30, 25, 28, 31,
 33, 27
]
```

- **Data Sorting**

```
ages = [
 22, 22, 23, 24, 24, 25, 25, 25, 25, 26,
 26, 26, 27, 27, 27, 28, 28, 28, 28, 28,
 29, 29, 30, 30, 30, 31, 31, 31, 32, 33,
 34, 34
]
```

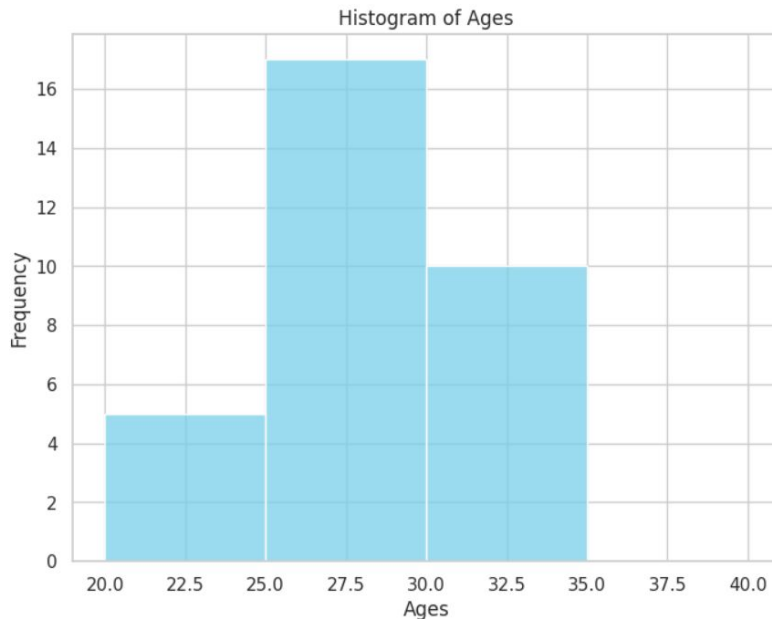
# Frequency Distribution (Numerical Data)

- **Create Intervals:** *Divide data into intervals or bins.*
  - Bins: [(20, 24), (25, 29), (30, 34), (35, 39)]
- **Count Frequencies:** *Count how many ages fall into each interval.*
- **Create Frequency Table:** *Organize the counts into a table.*

| Interval | Frequency |
|----------|-----------|
| 20 - 24  | 5         |
| 25 - 29  | 17        |
| 30 - 34  | 10        |
| 35 - 39  | 0         |

# Frequency Distribution (Numerical Data)

- **Visual Representation via Histogram**



- Histogram helps to visualize distributions of our data.
- Distributions means how your data look like



# PERCENTILES & OUTLIERS

# Percentiles & Outliers

- Percentile is a value below which a certain percentage of observation lie.
- Interpretation:
  - If student scored in the 75% percentile, they performed better than 75% of the students.

$$\text{Percentile rank of } x = \frac{\text{\# of value below } x}{n}$$

***75% percentile means 75% of data points are below the value at 75% percentile***

# Percentiles (Calculations)

- Consider data points:

```
data = [2, 2, 3, 4, 5, 5, 5, 6, 7,
8, 8, 8, 8, 8, 9, 9, 10,
11, 11, 12]
```

- Sort in ascending order

```
data = [2, 2, 3, 4, 5, 5, 5, 6, 7,
8, 8, 8, 8, 8, 9, 9, 10,
11, 11, 12]
```

- From the above data, what is the percentile ranking of 10?**
  - Percentile ranking of 10 =  $(16 / 20) * 100$   
= 80%
  - This means 80% of the entire distributions is less than 10.



# Percentiles (Calculations)

- From the data points, What value exists at percentile ranking of 25%?

$$\text{Value} = \frac{\text{Percentile}}{100} * (n + 1)$$

$$\begin{aligned}\text{Value} &= (25 / 100) * (20 + 1) \\ &= 5.25 \text{ (index position)}\end{aligned}$$

- Here, 5.25 is in between 5th and 6th index,
- 5th index value = 5
- 6th index value = 5
- Value =  $(5 + 5) / 2 = 5$

```
import numpy as np
```

```
25% percentile
```

```
data = [2, 2, 3, 4, 5, 5, 5, 6, 7,
8, 8, 8, 8, 8, 9, 9, 10,
11, 11, 12]
```

```
value = np.percentile(data, 25)
```

```
print(value)
```

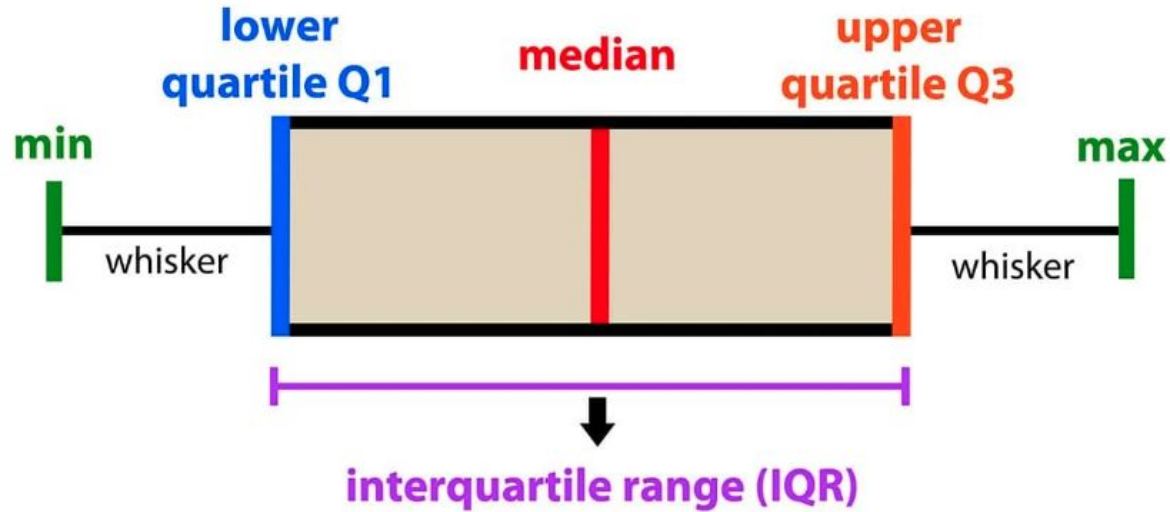
```
Output: 5.0
```

# Five Number Summary

- Minimum
- First Quartile (Q1) → 25% percentile
- Median (Q2) → 50% percentile
- Third Quartile (Q3) → 75% percentile
- Maximum

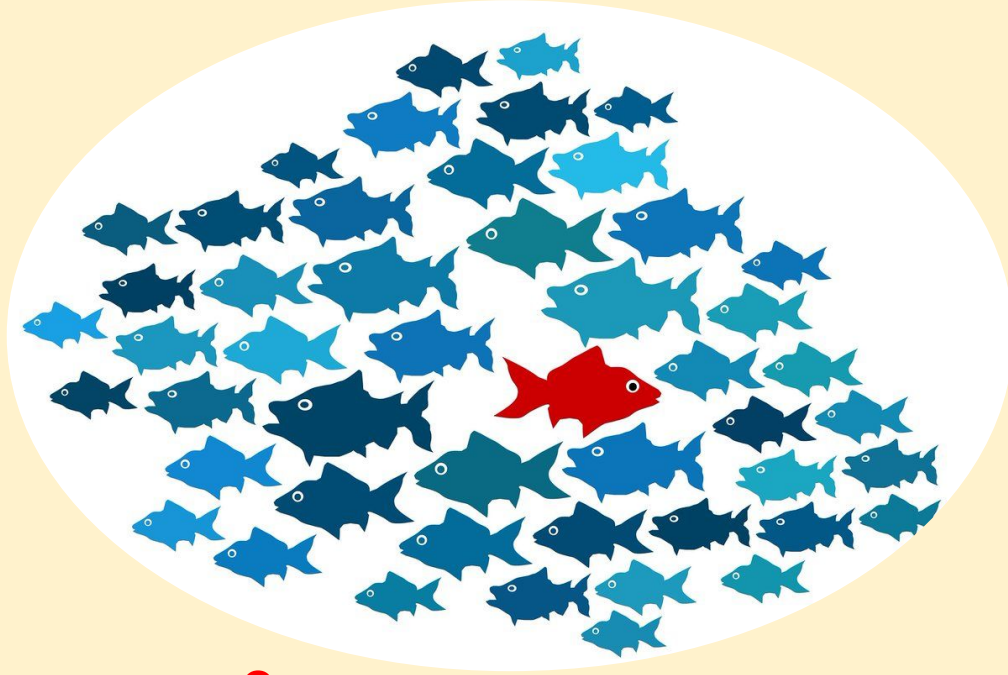
$$\text{IQR} = Q3 - Q1$$

# Box plot



***“Box plot can be used to determine outliers”***

```
sns.boxplot(data=data)
```



**“Outliers Detection”  
(IQR)**

# Outliers Detection Using IQR (Steps)

1. **Sort the Dataset in ascending order**
2. **Calculate Q1 and Q3**
  - Q1: 25% percentile
  - Q3: 75% percentile
3. **Compute Interquartile Range (IQR)**
  - IQR:  $Q3 - Q1$
4. **Compute Lower Fence and Upper Fence**
  - Lower Fence:  $Q1 - 1.5 * IQR$
  - Upper Fence:  $Q3 + 1.5 * IQR$
5. **Make Conclusion**
  - Data points below Lower Fence and,
  - Data points above Upper Fence are Outliers.

# Outliers Detection Using IQR (Lab)

```
dataset = [11, 10, 12, 14, 12, 15, 14, 13, 15, 102,
 12, 14, 17, 19, 107, 10, 13, 12, 14, 12,
 108, 12, 11, 14, 13, 15, 10, 15, 12, 10,
 14, 13, 15, 10]
```

```
Step 1: sort the dataset in ascending order
```

```
dataset = sorted(dataset)
```

```
print(dataset)
```

```
Step 2: calculate Q1 and Q3
```

```
Q1: 25% percentile
```

```
Q3: 75% percentile
```

```
Q1, Q3 = np.percentile(dataset, [25, 75])
```

```
print(f"Q1: {Q1}")
```

```
print(f"Q3: {Q3}")
```

```
Step 3: Compute IQR
```

```
iqr = Q3 - Q1
```

```
print(f"IQR: {iqr}")
```

```
Step 4: Compute Lower Fence and Upper Fence
```

```
lower_fence = Q1 - (1.5 * iqr)
```

```
upper_fence = Q3 + (1.5 * iqr)
```

```
print(f"Lower Fence: {lower_fence}")
```

```
print(f"Upper Fence: {upper_fence}")
```

```
Step 5: Make Conclusion
```

```
def detect_outlier_using_iqr(data, iqr,
lower_fence, upper_fence):
```

```
 outliers = []
```


```
 for i in data:
```

```
 if i < lower_fence or i > upper_fence:
```

```
 outliers.append(i)
```

```
 return outliers
```

# Notation (Population vs Sample)

| Parameter          | Population Notation                                                               | Sample Notation | Descriptions |
|--------------------|-----------------------------------------------------------------------------------|-----------------|--------------|
| Mean               |  |                 |              |
| Variance           |                                                                                   |                 |              |
| Standard Deviation |                                                                                   |                 |              |
| Size / Count       |                                                                                   |                 |              |
| Data point         |                                                                                   |                 |              |
| Sum                |                                                                                   |                 |              |